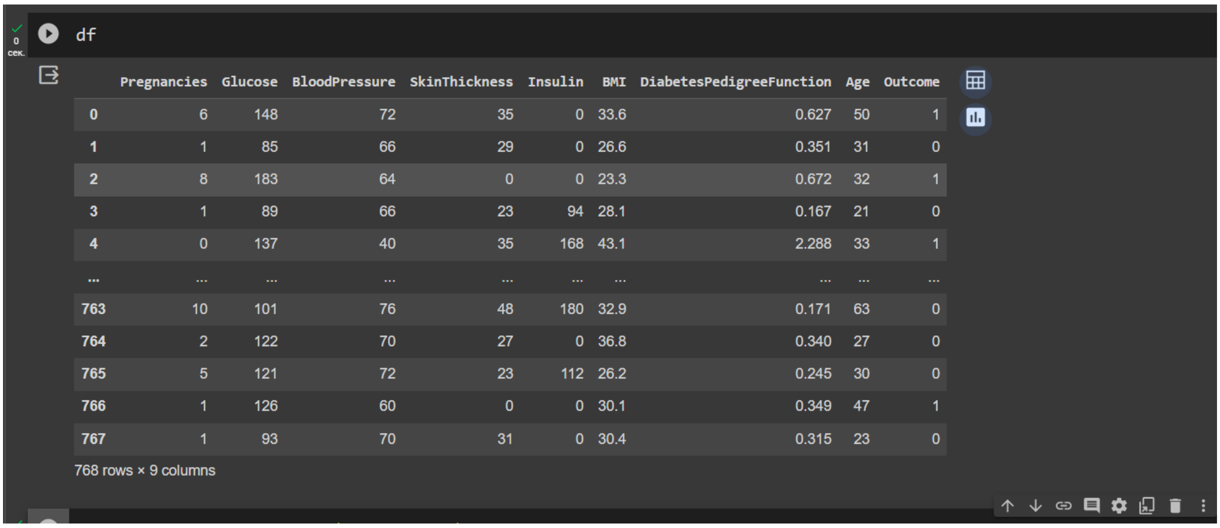
Diabetes Data Analysis

Diabetes is a chronic disease that affects how your body processes blood sugar

Diabetes mellitus, more commonly known simply as diabetes, refers to a group of diseases that can cause high levels of glucose (a type of sugar) in your blood.

Diabetes happens when your pancreas can't produce enough of the hormone insulin or your body becomes resistant to it.

Symptoms of diabetes are feeling tired, hungry, or excessively thirsty, and passing more urine (wee) than usual.



```
df
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

For analysis, let's take a data set: https://www.kaggle.com/datasets/aemyjutt/diabetesdataanslysis/data

Pregnancies: which person is count time pregnent

Glucose: level of sugar

BloodPressure: blood levels stable of not

SkinThickness: tells about your body skin

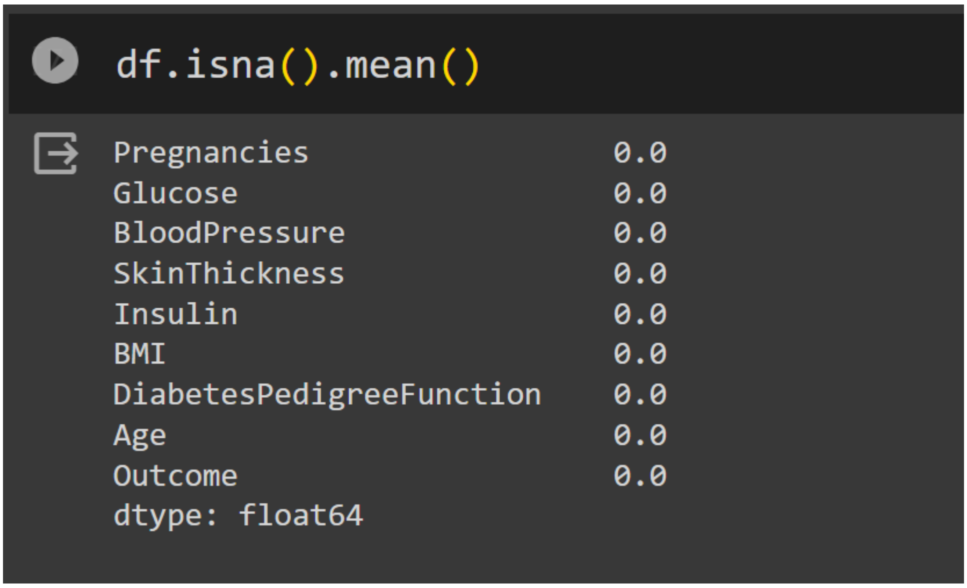
Insulin: need or not

BMI:tests

DiabetesPedigreeFunction: more info

Age:adult or older**Outcome:**result 1-bad,0-good

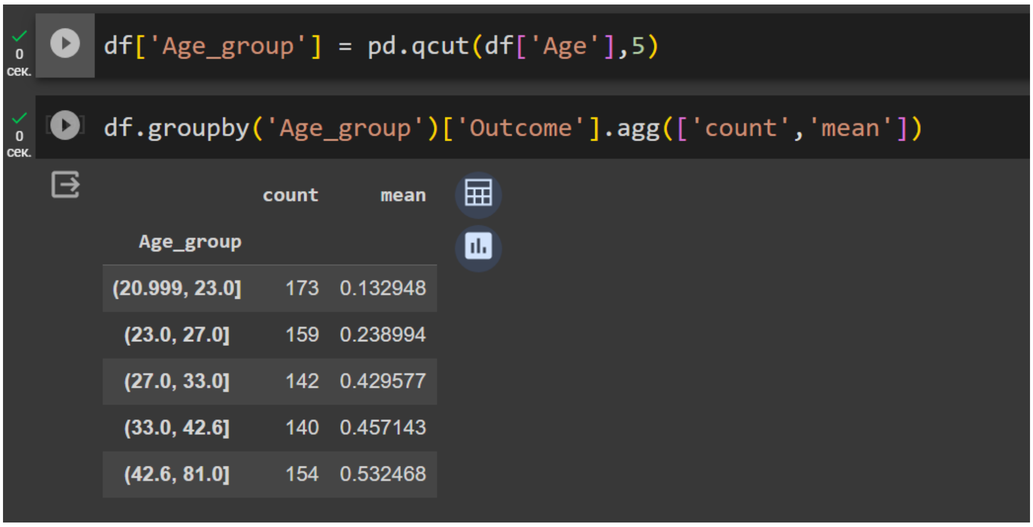
First, let’s check if there are any gaps in the data set.



```
df.isna().mean()
```

```
Pregnancies                 0.0
Glucose                     0.0
BloodPressure               0.0
SkinThickness               0.0
Insulin                     0.0
BMI                         0.0
DiabetesPedigreeFunction    0.0
Age                         0.0
Outcome                     0.0
dtype: float64
```

Let's take two indicators: **age** and the **result of diabetes**
In order to better read the data and see patterns, let’s take and divide the age into 5 groups and read the average value and count .In order to be able to loosen the data

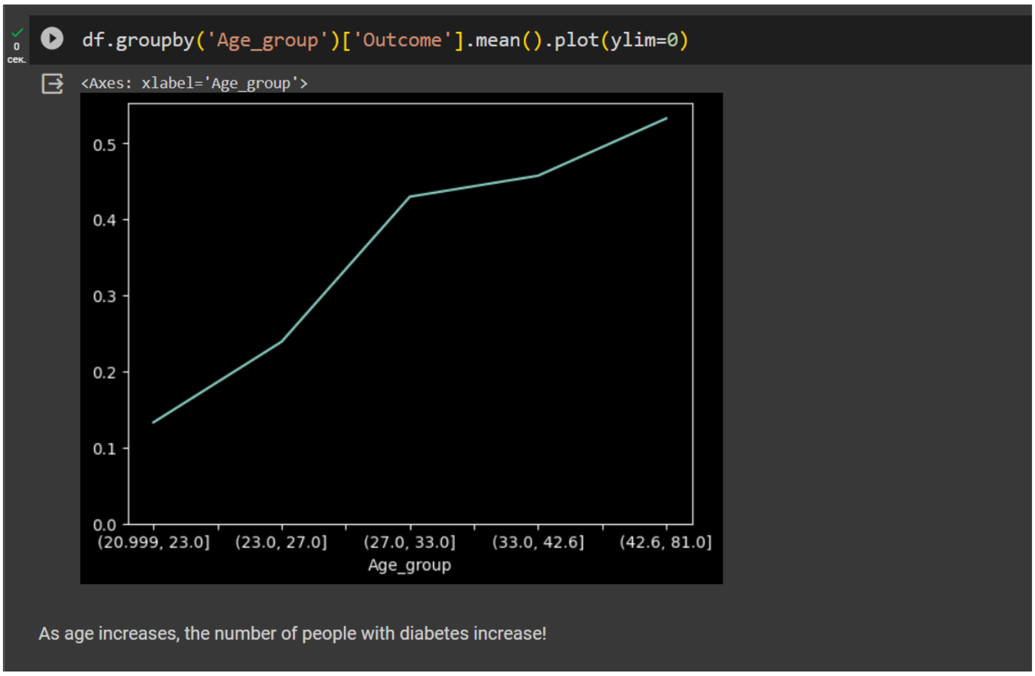


```
df['Age_group'] = pd.qcut(df['Age'],5)
```

```
df.groupby('Age_group')['Outcome'].agg(['count','mean'])
```

Age_group	count	mean
(20.999, 23.0]	173	0.132948
(23.0, 27.0]	159	0.238994
(27.0, 33.0]	142	0.429577
(33.0, 42.6]	140	0.457143
(42.6, 81.0]	154	0.532468

Let's build a graph and see the result.The results show that after 27 ages, the number of people with diabetes increases

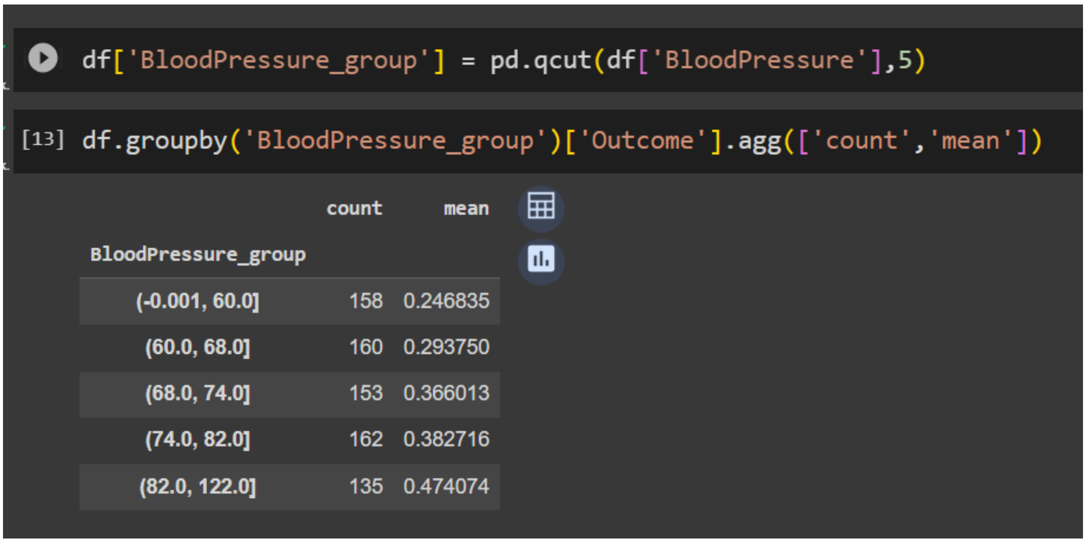


```
df.groupby('Age_group')['Outcome'].mean().plot(ylim=0)
```

```
<Axes: xlabel='Age_group'>
```

As age increases, the number of people with diabetes increase!

Let's look at the following indicator **BloodPressure** and the connection between diabetes

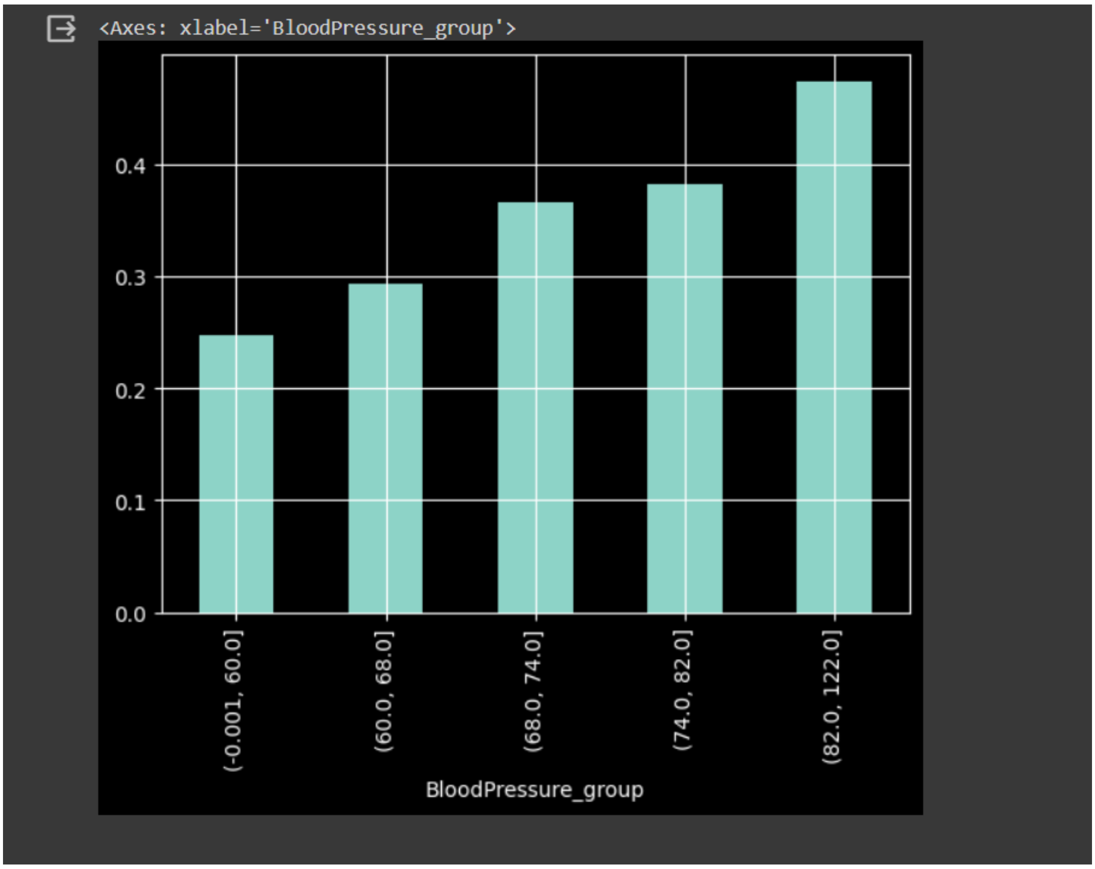


```
df['BloodPressure_group'] = pd.qcut(df['BloodPressure'],5)
```

```
[13] df.groupby('BloodPressure_group')['Outcome'].agg(['count','mean'])
```

BloodPressure_group	count	mean
(-0.001, 60.0]	158	0.246835
(60.0, 68.0]	160	0.293750
(68.0, 74.0]	153	0.366013
(74.0, 82.0]	162	0.382716
(82.0, 122.0]	135	0.474074

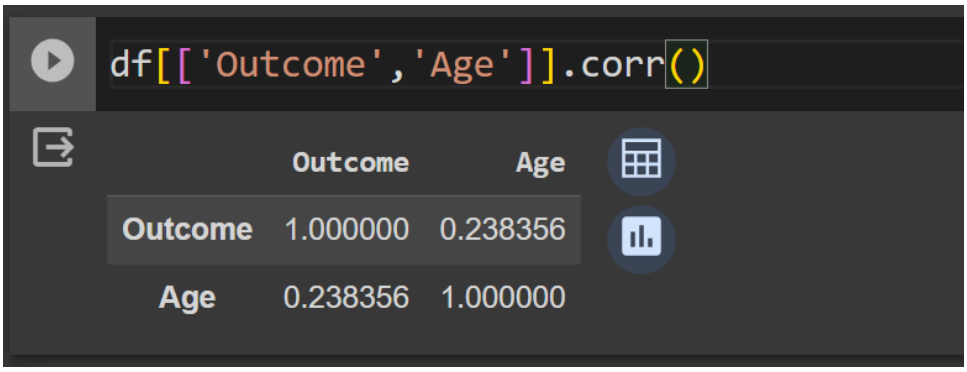
Let's divide the pressure level into groups for better analysis and build a graph.



```
<Axes: xlabel='BloodPressure_group'>
```

From the graph we see that diabetes is more common in people with high blood pressure, which means there is a connection between these indicators and people who have high blood pressure should check their level sugar in body.

Let's look for correlation between data

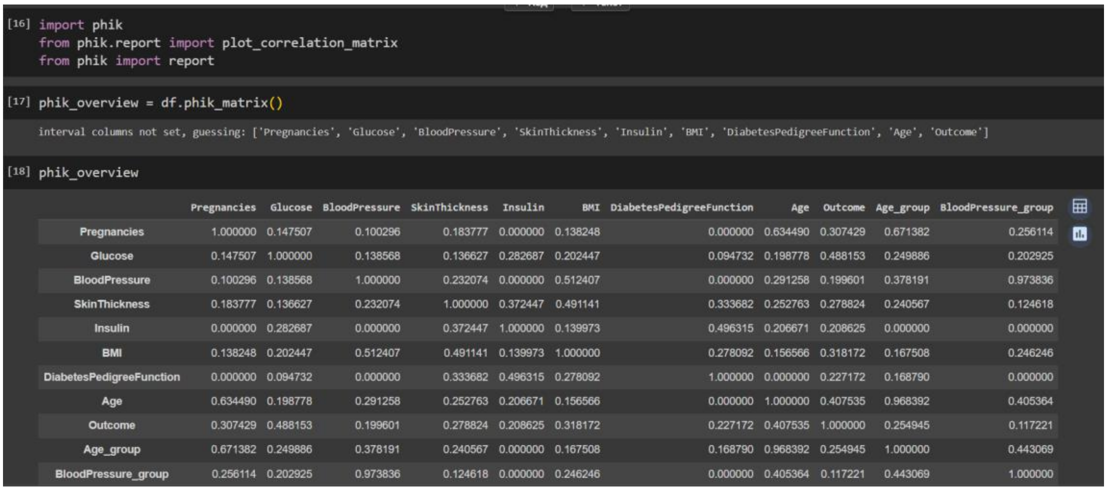


```
df[['Outcome','Age']].corr()
```

	Outcome	Age
Outcome	1.000000	0.238356
Age	0.238356	1.000000

As we see the correlation is about 20%, this is not enough for intelligence analysis. Here we do not see linear connections.Let's try to find connections through Phik (ϕk).

Phik (ϕk) is a new and practical correlation coefficient that works consistently between categorical, ordinal and interval variables, captures non-linear dependency and reverts to the Pearson correlation coefficient in case of a bivariate normal input distribution.



```
[16] import phik
     from phik.report import plot_correlation_matrix
     from phik import report
```

```
[17] phik_overview = df.phik_matrix()
```

```
interval columns not set, guessing: ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome']
```

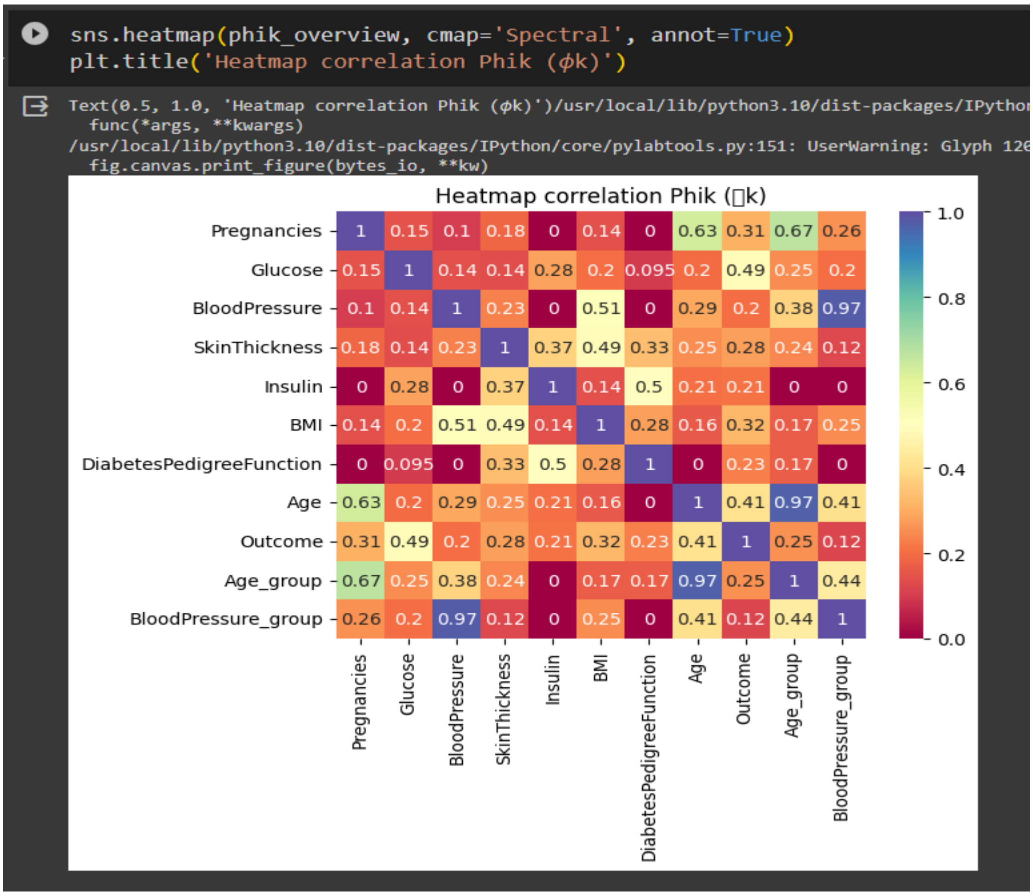
```
[18] phik_overview
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	Age_group	BloodPressure_group
Pregnancies	1.000000	0.147507	0.100296	0.183777	0.000000	0.138248	0.000000	0.634490	0.307429	0.671382	0.256114
Glucose	0.147507	1.000000	0.138568	0.136627	0.282687	0.202447	0.094732	0.198778	0.488153	0.249886	0.202925
BloodPressure	0.100296	0.138568	1.000000	0.232074	0.000000	0.512407	0.000000	0.291258	0.199601	0.378191	0.973836
SkinThickness	0.183777	0.136627	0.232074	1.000000	0.372447	0.491141	0.333682	0.252763	0.278824	0.240567	0.124618
Insulin	0.000000	0.282687	0.000000	0.372447	1.000000	0.139973	0.496315	0.206671	0.208625	0.000000	0.000000
BMI	0.138248	0.202447	0.512407	0.491141	0.139973	1.000000	0.278092	0.156566	0.318172	0.167508	0.246246
DiabetesPedigreeFunction	0.000000	0.094732	0.000000	0.333682	0.496315	0.278092	1.000000	0.000000	0.227172	0.168790	0.000000
Age	0.634490	0.198778	0.291258	0.252763	0.206671	0.156566	0.000000	1.000000	0.407535	0.968392	0.405364
Outcome	0.307429	0.488153	0.199601	0.278824	0.208625	0.318172	0.227172	0.407535	1.000000	0.254945	0.117221
Age_group	0.671382	0.249886	0.378191	0.240567	0.000000	0.167508	0.168790	0.968392	0.254945	1.000000	0.443069
BloodPressure_group	0.256114	0.202925	0.973836	0.124618	0.000000	0.246246	0.000000	0.405364	0.117221	0.443069	1.000000

A short introduction to ϕk

In many fields (not only data science), Pearson’s correlation coefficient is a standard approach of measuring correlation between two variables. However, it has some drawbacks:

- it works only with continuous variables,
- it only accounts for a linear relationship between variables,
- it is sensitive to outliers.
- The most similar metric to ϕk is Cramer’s ϕ, which is a correlation coefficient meant for two categorical variables and is also based on Pearson’s χ2 test statistic.



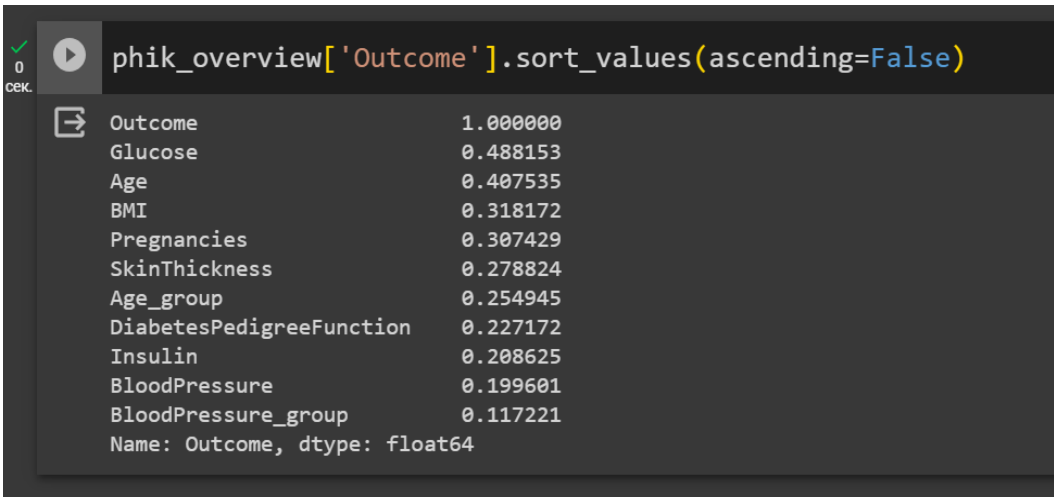
```
sns.heatmap(phik_overview, cmap='Spectral', annot=True)
plt.title('Heatmap correlation Phik (φk)')
```

```
Text(0.5, 1.0, 'Heatmap correlation Phik (φk)')/usr/local/lib/python3.10/dist-packages/IPytho
  func(*args, **kwargs)
/usr/local/lib/python3.10/dist-packages/IPython/core/pylabtools.py:151: UserWarning: Glyph 12
  fig.canvas.print_figure(bytes_io, **kw)
```

Heatmap correlation Phik (□k)

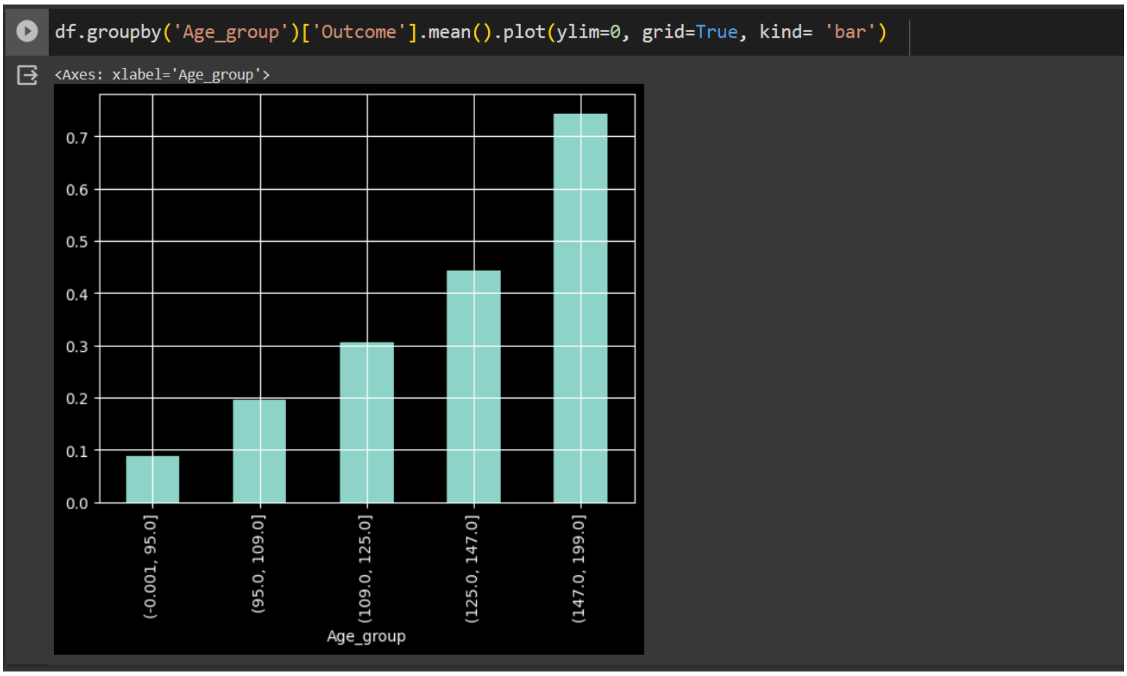
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	Age_group	BloodPressure_group
Pregnancies	1	0.15	0.1	0.18	0	0.14	0	0.63	0.31	0.67	0.26
Glucose	0.15	1	0.14	0.14	0.28	0.2	0.095	0.2	0.49	0.25	0.2
BloodPressure	0.1	0.14	1	0.23	0	0.51	0	0.29	0.2	0.38	0.97
SkinThickness	0.18	0.14	0.23	1	0.37	0.49	0.33	0.25	0.28	0.24	0.12
Insulin	0	0.28	0	0.37	1	0.14	0.5	0.21	0.21	0	0
BMI	0.14	0.2	0.51	0.49	0.14	1	0.28	0.16	0.32	0.17	0.25
DiabetesPedigreeFunction	0	0.095	0	0.33	0.5	0.28	1	0	0.23	0.17	0
Age	0.63	0.2	0.29	0.25	0.21	0.16	0	1	0.41	0.97	0.41
Outcome	0.31	0.49	0.2	0.28	0.21	0.32	0.23	0.41	1	0.25	0.12
Age_group	0.67	0.25	0.38	0.24	0	0.17	0.17	0.97	0.25	1	0.44
BloodPressure_group	0.26	0.2	0.97	0.12	0	0.25	0	0.41	0.12	0.44	1

On the graph we see the connection between glucose and diabetes, that is, the higher the level of glucose in the blood, the greater the chance of diabetes, and pregnant women are also susceptible to diabetes,but it's not a linear relationship,should not be trusted, you should always check the dependency



```
phik_overview['Outcome'].sort_values(ascending=False)
```

```
Outcome                     1.000000
Glucose                     0.488153
Age                         0.407535
BMI                         0.318172
Pregnancies                 0.307429
SkinThickness               0.278824
Age_group                   0.254945
DiabetesPedigreeFunction    0.227172
Insulin                     0.208625
BloodPressure               0.199601
BloodPressure_group         0.117221
Name: Outcome, dtype: float64
```



```
df.groupby('Age_group')['Outcome'].mean().plot(ylim=0, grid=True, kind= 'bar')
```

```
<Axes: xlabel='Age_group'>
```

We see that the relationship between age is highly correlated (it is not a linear relationship) As a result, age affects diabetes, that is, with age, the number of people with diabetes increases, and the presence of high blood pressure may indicate the presence of diabetes.With a linear correlation, the dependence on age was about 20%, with a Phik correlation the correlation was 40%.